Stationary Policies and Markov Policies in Borel Dynamic Programming

by

Manfred Schal* and William Sudderth**
Universität Bonn University of Minnesota

Technical Report No. 435

Abstract

The question of the existence of good Markov [good stationary] policies is studied for a general class of Borel [stationary] dynamic programming models. It is shown, for example, that Markov [stationary] policies are uniformly adequate if every transition law is absolutely continuous with respect to a fixed measure [and the reward function is positive or the model satisfies certain compactness and continuity conditions].

Key words and phrases: Dynamic programming, gambling, Markov policy, stationary policy, persistently optimal.

* Research supported by 'Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 72'

** Research supported by National Science Foundation Grant MCS 8100789

§1 Introduction

The elements of the model (S,A,D,q,r) are defined as follows:

(i) The state space S is a standard Borel space (that is, a nonempty Borel subset of some Polish space endowed with the σ-algebra of Borel subsets of S).

(ii) The action space A is also a standard Borel space.

(iii) D is a mapping which assigns to each $s \in S$ the set of available actions $D(s)$ which is a nonempty measurable subset of A . It is assumed that the set

$$\text{graph } D = \{(s,a); a \in D(s)\}$$

is a product measurable subset of $S \times A$ and contains the graph of a measurable map of S into A .

(iv) q is the law of motion (or transition law) and is a transition probability from graph D to S ; that is, q is a measurable mapping from graph D into the set $\mathbb{P}(S)$ of probability measures on S (equipped with the $*$-σ-algebra).

(v) The reward function r is a measurable mapping from graph $D \times S$ to $[-\infty,\infty)$.

The model does not explicitly contain a discount factor. A possible discount factor can be looked upon as the probability of staying in the system at each stage, that is not going to an extra absorbing state (cp. Federgruen et al. (1979)).
Throughout the paper 'measurable' means 'Borel measurable' unless a longer phrase such as 'universally measurable' or 'upper semianalytic' is used.
The history spaces are defined as $H_0 = S$, $H_{n+1} = \text{graph } D \times H_n$, $H = X_0^{\infty} \text{graph } D$. As usual, a randomized policy $\delta = (\delta_n)$ is defined as a sequence of transition probabilities from H_n to A such that $\delta_n(s_0,a_0,\ldots,s_n)$ assigns probability one to $D(s_n)$. The class of all randomized policies is denoted by Δ .
The class of all stationary policies is identified with the class $\mathbb{F}$ of measurable functions f from S to A such that $f(s) \in D(s)$. Finally a Markov policy $\delta \in \Delta_M$ is a sequence (f_n) where $f_n \in \mathbb{F}$.
Let ξ_n and α_n be the projections from H to the n-th state

space and action space, respectively, and define ξ_∞ as a constant value not contained in S.
Write

$$r_n = r(\xi_n, \alpha_n, \xi_{n+1}) .$$

Further $P_{s\delta}$ and $E_{s\delta}$ denote the probability measures and corresponding expectations for a given initial state s and a policy δ. If there is given an initial distribution λ on S, then $P_{\lambda\delta}$ is written for the measure

$$P_{\lambda\delta}(\cdot) = \int P_{s\delta}(\cdot)\lambda(ds) . \tag{1.0}$$

Define the <u>expected reward function</u> for δ by

$$I(\delta)(s) = I(\delta,s) = E_{s\delta}\left[\sum_{n=0}^{\infty} r_n\right],$$

and similarly define $I_+(\delta)$ by replacing r with $r^+ = r \vee 0$.
The <u>value function</u> is

$$v(s) = \sup\left\{I(\delta,s), \delta \in \Delta\right\} .$$

In order that the expected reward function will be well-defined for every δ and that the value function does not assume the value $+\infty$, the following integrability assumption is imposed in this section.
<u>Integrability assumption:</u> For all $s \in S$

$$v_+(s) = \sup\left\{I_+(\delta,s); \delta \in \Delta\right\} < +\infty .$$

Ornstein (1969) has shown that if the value function assumes the value $+\infty$ there need not be a good stationary policy even for positive countable models.
The appropriate notion of ε-optimality is not obvious for a general model. Blackwell (1965) has an example of a positive problem with countable state space in which no stationary policy achieves $v - \varepsilon$, while Ornstein (1969) has shown that, for all such models, some stationary policy earns at least $(1 - \varepsilon)v$.
This notion of multiplicative optimality also seems to be the

right one for positive gambling problems (Dubins and Sudderth (1979)). However, it is obviously not suitable for negative problems. Thus, what seems to be needed for general models is a notion corresponding to the multiplicative one where $v > 0$ and to the additive one where $v < 0$. The present formulation of such a notion is relative to a fixed nonnegative real (universally measurable) function b on the state space.
A policy δ is ε-optimal (relative to b) iff

$$I(\delta) \geq v - \varepsilon b .$$

Van der Wal (1981) has chosen b equal to v_+ in the context of stationary policies and b equal to $v_+ + 1$ in the context of Markov policies. In the present section the same choice of b will be made. However, in the next sections it will become clear that a rich class of candidates of b is available containing elements which may be much smaller then v_+ or $v_+ + 1$.
The collection $\mathbf{F}$ of stationary policies will be called locally adequate if

$$v = \sup\{I(f); f \in \mathbf{F}\} . \tag{1.1}$$

$\mathbf{F}$ has been proven locally adequate in a number of special cases including the positive Borel case (Blackwell, (1965) and a general Borel model with compactness and continuity assumptions (Schäl, 1983). In each of these cases, it is easy to modify the arguments so as to show that $\mathbf{F}$ is λ-nearly uniformly adequate in the sense that, for every $\varepsilon > 0$, there is some $f \in \mathbf{F}$ such that

$$\lambda\left[I(f) \geq v - \varepsilon b\right] \geq 1 - \varepsilon . \tag{1.2}$$

Here λ is a probability measure on S; v is universally measurable and even upper semianalytic (Bertsekas/Shreve(1978) Corollary 9.4.1). (The first result on universal measurability of v is in Stranch's (1966) paper; the notion of 'upper semianalytic' seems to have been introduced into the theory of dynamic programming in the paper by Blackwell et al. (1974).)
It seems to be more difficult to establish that $\mathbf{F}$ is uniformly adequate (relative to b) in the sense that, for every $\varepsilon > 0$,

there is some $f \in \mathbf{F}$ such that f is ε-optimal (relative to b where f may be universally measurable), that is

$$I(f) \geq v - \varepsilon b \ . \tag{1.3}$$

This follows for positive models with a countable state space from the result by Ornstein (1969) and for negative models with continuity and compactness assumptions from Schäl (1975). The question remains open for positive, Borel models and for the general model of Schäl (1983) under the so-called compactness and continuity assumptions (W) . However, Ornstein (1969) has a counter example in which the model is positive and all transition distributions have finite support. The situation is similar for positive, leavable, measurable gambling problems as is explained by Dubins and Sudderth (1979). They were, however, able to prove the uniform adequacy of stationary policies for absolutely continuous gambling problems. Similarly, we will give positive answers to the questions raised above under the assumption that the law of motion is absolutely continuous.
We find it useful to introduce a notion weaker than uniform adequacy but stronger than nearly uniform adequacy. Notice first that (1.2) can be rewritten as

$$P_{\lambda f}\left[I(f,\xi_o) \geq v(\xi_o) - \varepsilon b(\xi_o)\right] \geq 1 - \varepsilon \ . \tag{1.4}$$

$\mathbf{F}$ will be called λ-<u>persistently adequate</u> if, for every $\varepsilon > 0$, there is some $f \in \mathbf{F}$ such that

$$P_{\lambda f}\left[I(f,\xi_n) \geq v(\xi_n) - \varepsilon b(\xi_n) \text{ for all } n\right] = 1 \ . \tag{1.5}$$

Our main result, stated roughly here and precisely in the next section, is that for a very general class of stationary problems (containing both the dynamic programming problem and the gambling problem) λ-nearly uniform adequacy for all $\lambda \in \mathbb{P}(S)$ implies λ-persistent adequacy for all $\lambda \in \mathbb{P}(S)$ and in the absolutely continuous case even uniform adequacy. To be more precise, we must also assume nearly uniform adequacy for a certain class of submodels of the type first introduced by Ornstein. Here we will present a few corollaries of the main result.

Corollary 1.1. If the model is positive, that is $r \geq 0$, then the class $\mathbf{F}$ of stationary policies is λ-persistently adequate (relative to $b = v$) given any $\lambda \in \mathbf{P}(S)$.
The Corollary extends a result by Frid (1970) which also holds for Markov games (v. Dawen (1984)).
To make the assumption of absolute continuity precise we introduce the following condition:

Condition A: There is a probability measure $\mu \in \mathbf{P}(S)$ such that, for every $(s,a) \in$ graph D, the measure $q(s,a;\cdot)$ is absolutely continuous with respect to μ.
We remind the reader that any σ-finite measure is dominated by a probability measure; hence Condition A is satisfied if $q(s,a;\cdot)$ is absolutely continuous with respect to a σ-finite measure μ for all $(s,a) \in$ graph D.
It is clear from the results by Bertsekas and Shreve (1978) that one has to replace $\mathbf{F}$ by the set $\mathbf{F}_u$ of universally measurable stationary policies in order to get uniform ε-optimality, unless further regularity properties of the model are known. Also Strauch (1966) showed that v is not necessarily Borel from which it follows easily that $\mathbf{F}$ can fail to be uniformly adequate.

Corollary 1.2. If the model is positive and Condition A holds, then the class $\mathbf{F}_u$ is uniformly adequate (relative to $b = v$).

The next condition combines the absolute continuity assumption and the compactness and continuity assumptions.

Condition AC: 1) $D(s)$ is a compact subset of A for every $s \in S$,

2) Condition A is satisfied and the density functions $l(s,a,s') = q(s,a,ds')/\mu(ds')$ are jointly measurable in (s,a,s') and continuous in a.

3) $r(s,a,s')$ is upper semi-continuous in a for all $s,s' \in S$.

Corollary 1.3. Under the condition AC the class $\mathbf{F}$ is uniformly adequate (relative to $b = v_+$).

Finally an application to the persistent adequacy of the class Δ_M of Markov policies is given. Markov policies correspond to the stationary policies in a modified model with augmented state space such that the time parameter is incorporated in the state of the system. Hence results on stationary policies also apply

to Markov policies.

Corollary 1.4. The class Δ_M of Markov policies is always λ-persistently adequate (relative to $b = v_+ + 1$) given any $\lambda \in \mathbb{P}(S)$.

Corollary 1.5. Under condition A, the class $\Delta_{M,u}$ of universally measurable Markov policies is uniformly adequate (relative to $b = v_+ + 1$).
The last two corollaries extend a result by van der Wal (1983) for countable state space problems to Borel problems.

§2 The general model

In this section, a more general model $M = (S,A,D,q,r,g)$ is considered. The elements S,A,D,q,r are as in section 1. In addition g is defined as follows:

(vi) The payoff function g is a measurable mapping from H to $[-\infty,\infty)$ such that, for every $h = (s_0,a_0,s_1,a_1,\ldots) \in H$, $g(h) = r(s_0,a_0,s_1) + g(s_1,a_1,\ldots)$.

There are two special cases which illustrate the generality of g . If $g(h) = \sum_0^\infty r(s_m,a_m,s_{m+1})$, then g is the usual dynamic programming payoff of section 1. If $r = 0$, then g is shift-invariant and may be chosen as the payoff function studied by Dubins and Sudderth (1977b). (The measure theoretic structure of gambling problems differs from that of the dynamic programming models studied here as was explained by Blackwell (1976)).

A more general case was studied by Bodewig (1979) which contains the dynamic programming payoff and the Dubins and Savage payoff as special cases. Bodewig studies the function $g(h) = \overline{\lim_n} \sum_0^n r_m$. Then $g(h) = \overline{\lim_n}\, u(s_n) - u(s_0)$ if $r(s,a,s') = u(s') - u(s)$ for some utility function u on S . It is good to remember this example when looking at Assumption 4(c).

The value function v is defined, for every $s \in S$, by the equation

$$v(s) = \sup\{I(\delta,s); \delta \in \Delta\} \quad \text{where} \quad I(\delta,s) = E_{s\delta} g \ .$$

If this is to make sense, all the expectations on the right must exist. An even stronger assumption is made which implies that $v < +\infty$. Recall that a universally measurable function w on S is called excessive iff

$$\int w(s')q(s,a,ds') \le w(s)$$

for all $(s,a) \in$ graph D where also the existence of the integrals is supposed.

Assumption 1. There exists a non negative, real, universally measurable, excessive function $\bar{b}$ which dominates v .

In the frame work of section 1, $\bar{b}$ can be chosen as v_+ .

If g equals $\overline{\lim_n} u(s_n)$, then one may choose $\bar{b}(s) = \sup_\delta E_{s\delta}\left[\overline{\lim_n} u^+(\xi_n)\right]$.

Here is another convergence assumption, which, like the first, is not very restrictive in practice.

Assumption 2. $E_{s\delta}\left[\sup_n \sum_0^n r_m\right]^+ < \infty$ for every $s \in S$, and $\delta \in \Delta$.

Some definitions are needed. A mapping $\tau : H \to \{0,1,\ldots,\infty\}$ is an incomplete stopping time if $[\tau \le n] \in \sigma(\xi_0,a_0,\ldots,\xi_n)$ for every $n = 0,1,\ldots$. Let T be the set of all incomplete stopping times.

Lemma 2.1. $E_{s\delta}\left[\bar{b}(\xi_\tau)1_{[\tau<\infty]}\right] \le \bar{b}(s)$, $s \in S$, for all $\delta \in \Delta$, $\tau \in T$.

The proof follows from the optional sampling theorem of Doob, Fatou's inequality and the nonnegativity of b which implies $\underline{\lim}\, \bar{b}(\xi_{\tau\wedge n}) \ge \bar{b}(\xi_\tau)1_{[\tau<\infty]}$.

From the Lemma it follows that the expectations in the following definition are well-defined.

$$V(s) = \sup_{\delta\in\Delta}\inf_n \sup_{\tau\in T,\tau\ge n} E_{s\delta}\left[v(\xi_\tau)1_{[\tau<\infty]}\right].$$

The function V can be looked upon as a value function of the Dubins and Savage type. This will become clearer from the following remarks.

First we note that V is non negative since $\tau = \infty$ is an element of T. Moreover it is easy to prove that V does not change if v is replaced by $v^+ = v \vee 0$ in the definition. With this modification, it is again easy to prove that then T can be replaced by the set of a,s finite or everywhere finite or even bounded stopping times. In many cases V will be excessive. If there are no measurability problems as in the countable state space case one can imitate the proof of Theorem 2.14.1 or 3.3.1 by Dubins and Savage (1965) to prove the excessivity. Furthermore, under additional integrability assumptions and perhaps even without further assumptions, it can be shown that

$$V(s) = \sup_{\delta\in\Delta} E_{s\delta}\left[\overline{\lim_n} v^+(\xi_n)\right]$$

(cp. Sudderth (1971), Engelbert and Engelbert (1979)).

The latter function can be shown excessive (and upper semianalytic) by standard arguments. Finally, by Lemma 2.1, it is clear

that $\bar{b} \geq V$.

Assumption 3. There exists a nonnegative, real, upper semianalytic, excessive function b which dominates V .

By the preceding remark it is clear that one can choose $b = \bar{b}$. But, in general, V is much smaller then v . These are cases, where v is unbounded and V vanishes identically. The function b is the function which we will use for the concept of ε-optimality whereas $\bar{b}$ was introduced to avoid integrability difficulties. From the preceding remarks, it follows that in many cases one can choose b equal to V . Obviously, one can always replace b by $b + 1$; that is, one can always choose b bounded away from zero. The latter property would make the analysis much simpler as the following assumptions will show. However, Ornstein chooses b equal to v in the positive case which may even vanish for some states of the system. Hence, in order to cover Ornstein's result one is forced to take into consideration the following set

$$S_b = \{s; b(s) = 0\} .$$

Assumption 4. One of the following conditions are satisfied:

(a) $\inf b > 0$;

(b) $I(f) \geq 0$ for all $f \in \mathbf{F}$;

(c) $\overline{\lim}_{n\to\infty} \sum_{m=0}^{n} r_m \leq g$ $P_{s,f}$-a.s. for all $s \in S$, $f \in \mathbf{F}$.

Next we will assume that $\mathbf{F}$ is λ-nearly uniformly adequate (relative to b). This assumption probably implies that:

(2.0) on S_b there exists an optimal stationary policy f_b .

The proof of (2.0) in special cases will be made easier by the facts that S_b is closed in the sense that

$$q(s,a,S_b) = 1 \quad \text{for all} \quad a \in D(s) , \; s \in S_b .$$

and that every policy δ is equalizing on S_b in the sense that

$$\overline{\lim_n} \; E_{s\delta} v(\xi_n) \leq 0$$

since $V \leq 0$ on S_b. In the positive case where $b = v$ every policy is optimal on S_b.

Assumption 5a. For every probability measure $\lambda \in \mathbb{P}(S)$ and for any $\varepsilon > 0$ there exists a stationary policy $f \in \mathbb{F}$ such that

$$\lambda\left[I(f) \geq v - \varepsilon b\right] \geq 1 - \varepsilon .$$

For countable state space models Assumption 5a may be replaced by the assumption (1.1) (cp. van Dawen and Schäl (1983a)). If moreover the integrability assumption of section 1 holds, then the following Assumption 5b is not necessary as was shown by van Dawen (1983).
One of Ornstein's ideas consists in modifying the original model M so that only one action is available on a large set. That is the reason for the following definition. The model $M' = (S,A,D',q,r,g)$ is a submodel of M associated with the stationary policy f and the measurable subset G if

$$D'(s) = \begin{cases} \{f(s)\} & , s \in G , \\ D(s) & , s \notin G . \end{cases}$$

Assumption 5b. Assumption 5a holds for every submodel of M.

Theorem 1. For every $\lambda \in \mathbb{P}(S)$, $\mathbb{F}$ is λ-persistently adequate relative to b.

Theorem 2. If in addition Condition A holds, then $\mathbb{F}_u$ is uniformly adequate relative to b. The class $\mathbb{F}$ is uniformly adequate if, in addition, for all $\varepsilon > 0$ there is some "εb-conserving" $f \in \mathbb{F}$, that is

$$\int q(s,f(s),ds')\left[r(s,f(s),s') + v(s')\right] \geq v(s) - \varepsilon b(s) , \quad s \in S .$$

We remark that for Theorem 2 where Condiditon A holds it is sufficient that Assumption 5 only holds for $\lambda = \mu$ provided that also (2.0) holds.

The proofs of the theorems will be given in the next section. Here it will be shown how the corollaries of section 1 will follow from the theorems.

Proof of Corollaries 1.1 and 1.2: We choose $\bar{b} = v$, then Assumption 1 is fulfilled. The excessivity of v follows from the optimality equation. Assumption 2 follows from the integrability assumption of section 1. We choose $b = \bar{b}$, then Assumption 3 is satisfied. Obviously, Assumption 4b holds. The proof of Assumption 5a is similar to the proofs of the Corollaries 1.3, 1.4, 1.5 below.
Also Assumption 5b follows because $D(s)$ was arbitrary. Now the Theorems 1 and 2 apply. □

Proof of Corollary 1.3: We choose $\bar{b} = v_+ = b$. Then the Assumptions 1,2,3 are fulfilled. Obviously, Assumption 4c holds. For $m = 1,2,\ldots$ let v^m and I^m be the value function and the expected reward function for the model $M^m = (S,A,D,q,r \wedge m)$. Then, as follows from the Proposition in Schäl (1983)

$$v^m \uparrow v \quad \text{as} \quad m \to \infty \,. \tag{2.1}$$

Next let v_γ^m and I_γ^m be the value function and the expected reward function for the model M_γ^m with reward function $\gamma(r \wedge m) + (1-\gamma)v_-$, where v_- is the value function of the imbedded negative problem and discount factor γ as in section 2 (ibidem).
Then it follows from the proof of the Theorem (ibidem) that

$$v_\gamma^m \to v^m \quad \text{as} \quad m \to \infty \,. \tag{2.2}$$

Further, since the Condition (S) (ibidem) is satisfied by the present Condition AC, it follows from Lemma 2.1 (ibidem) that for every $\gamma \in (0,1)$ there exists some $f_\gamma \in \mathbf{F}$ such that

$$I_\gamma^m(f_\gamma) = v_\gamma^m \,. \tag{2.3}$$

Finally, we know from Lemma 2.8 (ibidem) that

$$I^m(f_\gamma) \geq v_\gamma^m \,. \tag{2.4}$$

Now choose some $\lambda \in \mathbf{P}(S)$ such that first $\lambda(S_b) = 0$. Then it follows from (2.1) that there is some m_o such that

$$(2.5) \qquad \lambda\left[v^{m_o} > v - \varepsilon/2b\right] > 1 - \varepsilon/2 \ .$$

From (2.2) if follows that there is some γ_o such that

$$(2.6) \qquad \lambda\left[v_{\gamma_o}^{m_o} > v^{m_o} - \varepsilon/2b\right] > 1 - \varepsilon/2 \ .$$

From (2.4) one has

$$(2.7) \qquad I(f_{\gamma_o}) \geq I^{m_o}(f_{\gamma_o}) \geq v_{\gamma_o}^{m_o} \ .$$

From (2.5),(2.6),(2.7) one obtains the desired inequality

$$\lambda\left[I(f_{\gamma_o}) \geq v - \varepsilon b\right] \geq 1 - \varepsilon \ .$$

From (2.1),(2.2),(2.3) we conclude that v and similarly v_+ are measurable. Hence S_b is measurable. On, S_b we have

$$\int q(s,a,ds')r^+(s,a,s') = 0 \ , \ a \in D(s) \ , \ s \in S_b \ .$$

Hence we may assume without loss of generality that $r(s,a,s') \leq 0$, $s,s' \in S_b$, $a \in D(s)$, that is, we have a negative problem on S_b . Now the compactness and continuity assumptions imply that (2.0) holds (cp. Schäl (1975)). Now choose any $\lambda \in \mathbb{P}(S)$. Then choose some f' as in the first part for the conditional probability $\lambda(\cdot|S_b^c)$ and define

$$f = f' \ \text{ on } \ S_b^c \ \text{ and } \ f = f_b \ \text{ on } \ S_b \ .$$

Then it is easy to see that

$$\lambda\left[I(f) \geq v - \varepsilon b\right] \geq 1 - \varepsilon \ .$$

Assumption 5b follows from the same proof since any submodel satisfies the same compactness and continuity condition. Finally the existence of an εb-conserving $f \in \mathbb{F}$ will be proved. On S_b we may choose $f = f_b$. Now define for $m = 1,2,\ldots$

$$u_m(s,a) = \int \mu(ds')l(s,a,s')\left[\left(r(s,a,s') + v(s)\right) \wedge m\right]$$

then u_m is upper semicontinuous in a by Fatou's Lemma on the compact set $D(s)$ and of course measurable on graph (D) . Now, by a known selection theorem, cp. Brown and Purves (1973), there exist $f_m \in F$ such that

$$u_m(s,f_m(s)) = \max_{a\in D(s)} u_m(s,a) = u_m^*(s) \ , \ s \in S \ .$$

Now $u_m \uparrow u$, say, and hence

$$u_m^*(s) \uparrow u^*(s) = \sup_{a\in D(s)} u(s,a) \ , \quad s \in S \ .$$

Define

$$B_m = \left[u_m^* > u^* - \varepsilon b \ , \ u_k^* \le u^* - \varepsilon b \ , \ k < m\right]$$

and

$$f = f_m \quad \text{on} \quad B_m \ ,$$

then

$$\int B_m \supset S_b^c$$

and

$$u(s,f(s)) \ge u^*(s) - \varepsilon b(s) \ , \ s \notin S_b \ . \quad \square$$

Proof of Corollaries 1.4 and 1.5: The proof will show that the corollaries are even true after obvious modifications, when the original model is a non-stationary Markovian model in sense of Schäl (1975, section 8). There it was pointed out that Markov policies correspond in an obvious way to stationary policies in a model $\hat{M}$ with angmented state space $\hat{S} = S \times \{0,1,2,\ldots\}$. The Integrability assumption of section 1 carries over to the model $\hat{M}$. Because of the stationarity of the original model we know for the value function of model $\hat{M}$: $\hat{v}(s,n) = v(s)$ and $\hat{\Diamond}_+(s,n) = v_+(s)$. We choose $\hat{\bar{b}} = \hat{b} = \hat{\Diamond}_+ + 1$. Then the Assumptions 1,2,3 are fulfilled. Obviously, Assumption 4a holds. Assumption 5 can be proved in the frame work of the original model by

replacing stationary policies by Markov policies. However, in order to show Assumption 5b, one has to consider also models where the sets of available actions depend on the time parameter. For $m = 1,2,\ldots$ let v^m and I^m be the value function and the objective function for the model M^m where r is replaced with $r \wedge m$. Then (2.1) holds. Next for $0 < \gamma < 1$ let v_γ^m and I_γ^m be the value function and the expected reward function for the model M_γ^m with reward function $\gamma^{n+1} r_n^+ \wedge m + r_n \wedge 0$. By substracting $\gamma^{n+1} m$ this model can be made a negative model. Here, the reward function for period n actually depends on n. One has

$$v_\gamma^m \uparrow v^m \quad \text{as} \quad \gamma \uparrow 1 . \tag{2.8}$$

For the model M_γ^m it follows from Stranch (1966, theorem 8.1) that given any $\lambda \in \mathbb{P}(S)$ there is some $\delta \in \Delta_M$ such that

$$\lambda\left[I_\gamma^m(\delta) \geq v - \varepsilon/3\right] = 1 . \tag{2.9}$$

Actually, Stranch's result for stationary models does not directly apply because we had to consider nonstationary models where the reward function and the set of available actions depend on the time parameter. However, one can imitate Strauch's result where his Theorem 4.3 can be replaced by Theorem 18.2 of Hinderer (1980).

Now $I(\delta) \geq I^m(\delta) \geq I_\gamma^m(\delta)$. The rest goes through as in the proof of Corollary 1.3 using $b \geq 1$. □

§3 Preliminary Lemmas

Operators of the sort introduced into dynamic programming by Blackwell will be used throughout the proofs. For every $\delta \in \Delta$, $\tau \in T$, $s \in S$ and universally measurable mapping $w : S \to [-\infty,\infty)$, let

$$(3.1) \qquad \begin{aligned} Q_\delta^\tau w(s) &= E_{s\delta}\left[w(\xi_\tau)1_{[\tau<\infty]}\right], \\ L_\delta^\tau w(s) &= E_{s\delta}\left[\left\{\sum_{m=0}^{\tau-1} r_m + w(\xi_\tau)\right\}1_{[\tau<\infty]} + g\ 1_{[\tau=\infty]}\right] \end{aligned}$$

whenever the integrals are well-defined.
Several Lemmas are needed. Proofs are omitted when they are obvious or straight forward.

Lemma 3.1. $L_\delta^\tau(w+w') = L_\delta^\tau w + Q_\delta^\tau w'$ if all these quantities are well defined.

It is useful to have a formula which decomposes the return from a policy $\delta \in \Delta$ into that earned before time $\tau \in T$ and that earned there after. To get such a formula, first write

$$(3.2) \qquad I(\delta,s) = E_{s\delta}\left[\left\{\sum_{m=0}^{\tau-1} r_m + g(\xi_\tau,\ldots)\right\}1_{[\tau<\infty]} + g\ 1_{[\tau=\infty]}\right].$$

Next, let $h = (s_0,a_0,\ldots)$ be a history in H and, for $n = 0,1,\ldots$, define $p_n = (s_0,a_0,\ldots,s_n)$. Let $\delta[p_n]$ be the δ-conditional policy given the partial history p_n; that is, $\delta[p_n]$ is the policy defined by

$$\delta[p_n]_k(s_n,a_0',\ldots,s_k') = \delta_{n+k}(s_0,a_0,\ldots,s_n,a_0',\ldots,s_k').$$

For $\tau \in T$, $\tau(h) < \infty$, let $\delta[p_\tau] = \delta[p_{\tau(h)}]$. It is straight forward to check that

$$I\left(\delta[\xi_0,\alpha_0,\ldots,\xi_\tau],\xi_\tau\right) = E_{s\delta}\left[g(\xi_\tau,\alpha_\tau,\ldots) \mid \xi_0,\alpha_0,\ldots,\xi_\tau\right]$$

with probability one under δ on the set $[\tau < \infty]$. Hence (3.2) can be written as

$$(3.3) \quad I(\delta,s) = E_{s\delta}\left[\left\{\sum_{m=0}^{\tau-1} r_m + I\left(\delta[\xi_0,\alpha_0,\ldots,\xi_\tau],\xi_\tau\right)\right\}1_{[\tau<\infty]} + g1_{[\tau=\infty]}\right]$$

This is similar to a formula of gambling theory (Dubins and Savage (1965), Dubins and Sudderth (1977a)). The next lemma is immediate from (3.1) and (3.3).

Lemma 3.2. If $\delta \in \Delta$ and $\tau \in T$, then $I(\delta) \le L_\delta^\tau v$ provided that $L_\delta^\tau v$ is well-defined.
Suppose now that $\delta, \delta' \in \Delta$ and $\tau \in T$. Define a new policy $\delta^* = \varphi(\delta, \delta', \tau)$ to agree with δ prior to time τ and then conditionally equal δ'; that is,

$$\begin{aligned} \delta_n^*(\xi_o, \alpha_o, \ldots, \xi_n) &= \delta_n(\xi_o, \alpha_o, \ldots, \xi_n) \quad \text{on} \quad [\tau \ge n] \\ &= \delta'_{n-\tau}(\xi_\tau, \ldots, \xi_n) \quad \text{on} \quad [\tau < n] . \end{aligned}$$

In particular, $\delta^*[p_\tau] = \delta'$. Apply (3.1) and (3.3) to get another Lemma.

Lemma 3.3. If $\delta, \delta' \in \Delta$, $\tau \in T$, and $\delta^* = \varphi(\delta, \delta', \tau)$ then

$$I(\delta^*) = L_\delta^\tau I(\delta') ,$$

Lemma 3.4. If $\delta \in \Delta$ and $\tau \in T$, then $L_\delta^\tau v$ exists and $L_\delta^\tau v \le v$.

Proof. First it can be shown by known techniques (cp. Strauch (66) theorem 8.1, Hinderer (80) theorem 14.1) that for any $\lambda \in \mathbb{P}(S)$ and any $\varepsilon > 0$ there is some $\delta' \in \Delta$ such that

$$\lambda\left[I(\delta') \ge v - \varepsilon\right] = 1 .$$

If $\tau = \infty$ $P_{s\delta}$-almost surely, then $(L_\delta^\tau v)(s) = I(\delta, s) \le v(s)$. If not, choose $\lambda \in \mathbb{P}(S)$ such that $\lambda(B)$ equals $Q_\delta^\tau(1_B)(s)$ up to a normalizing constant and define $\delta^* = \varphi(\delta, \delta', \tau)$ as in Lemma 3.3. Then

$$L_\delta^\tau v \le L_\delta^\tau I(\delta') + \varepsilon = I(\delta^*) + \varepsilon \le v + \varepsilon \qquad \square$$

The following general form of the optimality equation is a consequence of Lemmas 3.2 and 3.4.

Lemma 3.5. If $\tau \in T$, then $v = \sup\{L_\delta^\tau v; \delta \in \Delta\}$.

For $n = 0,1,2,\ldots$ and $h = (s_0,a_0,s_1,\ldots) \in H$, define

$$\theta_n(h) = (s_n,a_n,s_{n+1},\ldots) .$$

Let $\tau \in T$ be defined only on $[\xi_0 \in E]$. Then let

$$\begin{aligned} \tau^0 &= 0 , \ \tau^1 = \tau \quad \text{on} \quad [\xi_0 \in E] \\ \tau^{n+1} &= k + \tau \circ \theta_k \quad \text{on} \quad \left[\tau^n = k, \xi_{\tau^m} \in E,\ m = 0,\ldots,n\right] \\ &= \infty \quad \text{on} \quad [\tau^n = \infty] \end{aligned}$$

Lemma 3.6. Let $f \in \mathbb{F}$, $\tau \in T$, and let $w : S \to [-\infty,\infty)$ be a universally measurable mapping dominated by v. Then $L_f^{\tau^n} w$ is a universally measurable mapping dominated by v and

$$L_f^{\tau^n} w = (L_f^\tau)^n w .$$

Proof. From Lemma 3.4 it follows that $L_f^{\tau^n} w$ is dominated by v and from Bertsekas and Shreve (1978, Proposition 7.46) then follows the universal measurability. From the Markov property we conclude that

$$E_{sf}\left[\left\{\sum_{k=t}^{\tau^m-1} r_k + w(\xi_{\tau^m})\right\} 1_{[\tau^m<\infty]} + g(\xi_{\tau^m},\alpha_{\tau^m},\ldots) 1_{[\tau^m=\infty]} \;\middle|\; \xi_0,\alpha_0,\ldots\xi_t\right]$$
$$= L_f^\tau w(\xi_t) \quad \text{on the set} \quad [\tau^{m-1} = t] .$$

Use this formula and induction to complete the proof. □

Lemma 3.7. Let $f \in \mathbb{F}$, $\tau \in T$, $\tau \geq 1$, $\Omega_\tau = \bigcup_n [\tau^n = \infty]$. Then

(a) $\overline{\lim_n} L_f^{\tau^n} 0(s) \leq E_{sf}\left[\left\{\overline{\lim_n} \sum_{m=0}^n r_m\right\} 1_{\Omega_\tau^c} + g\, 1_{\Omega_\tau}\right]$

(b) $\overline{\lim_n} L_f^{\tau^n} 0 \leq I(f)$ under Assumption 4b or 4c.

(Under Assumption 4a, the same inequality holds by (a) if one can prove that $P_{sf}(\Omega_\tau) = 1$.

Proof. From the monotone convergence theorem replacing g by g^+ and g^- we conclude

$$E_{sf} g\, 1_{[\tau^n=\infty]} \to E_{sf} g\, 1_{\Omega_\tau}$$

Defining

$$R_n = \sum_{m=0}^{\tau^n-1} r_m 1_{[\tau^n<\infty]}$$

we have

$$R_n \le \left(\sup_n \sum_{m=0}^{n} r_m\right)^+$$

and

$$\overline{\lim}\, R_n \le \left\{\overline{\lim}_n \sum_{m=0}^{n} r_m\right\} 1_{\Omega_\tau^c} \quad \text{since} \quad \tau^n \to \infty$$

Finally, by Fatou's Lemma using Assumption 2,

$$\overline{\lim}\, E_{sf} R_n \le E_{sf}\, \overline{\lim}\, R_n$$

and Part (a) follows. Part (b) is now clear under Assumption 4c. Under 4b one has

$$L_f^{\tau^n} 0 \le L_f^{\tau^n} I(f) = I(f) \quad \text{for all} \quad n$$

where the last equality is an instance of Lemma 3.3 with $\delta = \delta' = f$. □

We write $\hat{\mathbb{P}}(S)$ for the set of all defective probability measures λ on S (i.e., $\lambda(S) \le 1$).

Lemma 3.8. For $n = 1,2,\ldots$, $\varepsilon > 0$, $\lambda_m \in \hat{\mathbb{P}}(S)$, $m = 1,\ldots,n$ there is some $f \in \mathbb{F}$ such that

$$\lambda_m\Big(I(f) \ge v - \varepsilon b\Big) \ge (1-\varepsilon)\lambda_m(S) \quad m = 1,\ldots,n .$$

Proof. Assume without loss of generality that $\lambda_m(S) > 0$ for $m = 1,\ldots,n$. Choose f according to Assumption 5a such that

$$\frac{1}{n}\sum_{m=1}^{n} \lambda_m\Big(I(f) \ge v - \frac{\varepsilon}{n} b\Big) \Big/ \lambda_m(S) \ge 1 - \frac{\varepsilon}{n} . \qquad \square$$

§4 Proofs of Theorem 1 and 2

The proofs presented below uses ideas of Ornstein (1969). His ideas have also been used by Barbosa-Dantas (1966) and Frid (1970) in the theory of positive dynamic programming, and by Dubins and Sudderth (1979) and Sudderth (1969) in measurable gambling theory. Because we treat a general dynamic programming model (i.e. neither positive nor negative), we have had to modify the arguments of previous authors. However, our proof clearly owes much to them.

First we will turn to Ornstein's idea of modifying the model M so that only one action is available on a large set G.

For $\lambda \in \mathbf{P}(S)$, $f \in \mathbf{F}$, $0 < \eta < 1$, define the universally measurable set $G(f,\eta,M) = \left[I(f) \geq v - \eta b\right]$.

Choose $G_\lambda(f,\eta,M)$ such that $G_\lambda \subset G$, G_λ is (Borel-) measurable and $P_{\lambda f}(\xi_n \in G - G_\lambda) = 0$ for all n.

Now we can define a submodel $M' = M'(\lambda,f,\eta,M) = (S,A,D',p,r,g)$ where

$$D'(s) = \{f(s)\}, \quad s \in G_\lambda$$
$$= D(s), \quad s \notin G_\lambda$$

and a stopping time

$$\tau' = \inf\{n, \xi_n \notin G_\lambda\}$$

Further choose $N_\lambda \subset S$ such that N_λ is (Borel) measurable, $\lambda(N_\lambda) = 0$ and $P_{sf}(\xi_n \in G - G_\lambda) = 0$ for all n, $s \notin N_\lambda$, (cp. (1.0)). Then define

$$E_\lambda(f,\eta,M) = \left(G_\lambda(f,\eta^2,M) - N_\lambda\right) \cap G_\lambda(f,\eta,M).$$

As a consequence we have

(4.1) $\quad P_{sf}(\xi_\tau \in G - G_\lambda) = 0$ for all $\tau \in T$, $s \notin N_\lambda$.

Lemma 4.1. (a) M' satisfies all the assumptions made on M

(b) $E_\lambda \subset G_\lambda \subset G$, $E_\lambda \subset G_\lambda(f,\eta^2,M)$,
$\lambda(G_\lambda(f,\eta^2,M) - E_\lambda) = 0$

(c) $Q_f^{\tau'} b \leq \eta b$ on E_λ,

(d) $v' \geq v - \eta b$ where v' is the value function for M' .

Proof. a) It is easy to see that the model M' inherits the Assumptions 1,2,3,4. Assumption 5 is satisfied because any submodel of M' is a submodel of M .
b) These statements are obvious from the definitions.
c) Let $s \in E_\lambda$, then

$$\begin{aligned} v(s) - \eta^2 b(s) &\leq I(f,s) \\ &= L_f^{\tau'} I(f)(s) \quad \text{by Lemma 3.3} \\ &\leq L_f^{\tau'}(v - \eta b)(s) \quad \text{by (4.1)} \\ &= L_f^{\tau'} v(s) - \eta Q_f^{\tau'} b(s) \quad \text{by Lemma 3.1} \\ &\leq v(s) - \eta Q_f^{\tau'} b(s) \quad \text{by Lemma 3.4} \end{aligned}$$

d) First let $s \in G_\lambda$, then $v'(s) \geq I(f,s) \geq v(s) - \eta b(s)$. Now let $s \notin G_\lambda$ and $\eta' > 0$. Choose $\delta \in \Delta$ such that $I(\delta,s) \geq v(s) - \eta'$ and define $\sigma = \inf\{n, \xi_n \in G_\lambda\}$ and $\pi = \varphi(\delta,f,\sigma)$ according to Lemma 3.3. Then $\pi \in \Delta'$ and

$$\begin{aligned} v'(s) &\geq I(\pi,s) = L_\delta^\sigma I(f)(s) \geq L_\delta^\sigma (v - \eta b)(s) \\ &= L_\delta^\sigma v(s) - \eta Q_\delta^\sigma b(s) \\ &\geq I(\delta,s) - \eta b(s) \quad \text{by Lemma 3.2 and Lemma 2.1} \\ &\geq v(s) - \eta' - \eta b(s) \ . \quad \square \end{aligned}$$

Next we will construct for a given $\lambda \in \mathbb{P}(S)$ and $0 < \varepsilon < 1/2$ a policy $f \in \mathbb{F}$ and a stopping time $\tau \in T$. The policy f will be shown to satisfy (1.5) (with ε replaced by 4ε) thus proving Theorem 1. The construction is made by means of a sequence $f_1, f_2, \ldots$ of elements of $\mathbb{F}$, two sequences $G_1, G_2, \ldots$ and $E_1, E_2, \ldots$ of measurable subsets of S , and a sequence $\{\tau(k)\}$ in T satisfying

$$\tau(k) = \inf\{n, \xi_n \notin G_k\} \ , \tag{4.2}$$

$$E_k \subset G_k \ , \tag{4.3}$$

$$f_k = f_j \text{ on } G_j \text{ for } j \leq k \ . \tag{4.4}$$

The policy f and stopping time τ are then defined by

$$(4.5) \qquad \begin{aligned} f &= f_k \quad \text{on} \quad G_k \\ &= f_1 \quad \text{on} \quad S - UG_k \end{aligned}$$

$$(4.6) \qquad \begin{aligned} \tau &= \tau(k) \quad \text{on} \quad \left[\xi_o \in E_k - (E_1 U \ldots U E_{k-1})\right] \\ &= 1 \quad \text{on} \quad S - UE_k \ . \end{aligned}$$

At the k^{th} stage of the construction, we must consider, in addition to λ, other initial distributions arising from the distributions of $\xi_{\tau m}$. So we introduce the following notation for $m = 0,\ldots,k-1$,

$$(4.7) \qquad \begin{aligned} \lambda_{k0} &= \lambda \\ \lambda_{k1} &= P_{\lambda f_{k-1}}\left[\xi_o \in E_1 U \ldots U E_{k-1}, \xi_\tau \in \cdot\right] \\ &\ldots \\ \lambda_{km} &= P_{\lambda f_{k-1}}\left[\xi_o \in E_1 U \ldots U E_{k-1}, \ldots, \xi_{\tau m-1} \in E_1 U \ldots U E_{k-1}, \xi_{\tau m} \in \cdot\right] \end{aligned}$$

Now choose $\lambda_k \in \mathbb{P}(S)$ such that

$$(4.8) \qquad \lambda_{kj} << \lambda_k \quad \text{for} \quad j = 0,\ldots,m \ .$$

The next Lemma shows how the construction will be made.

<u>Lemma 4.2.</u> Let $0 < \varepsilon < 1/2$, $\eta_k = \varepsilon/2^k$, $\varepsilon_k = \eta_1 + \ldots + \eta_k$, $k = 1,2,\ldots$.
Then there exist $f_1, f_2, \ldots \in \mathbb{F}$ such that for

$$\begin{aligned} M_k &= M'(\lambda_k, f_k, \eta_k, M_{k-1}) \quad \text{where} \quad M_o = M \\ G_k &= G_{\lambda_k}(f_k, \eta_k, M_{k-1}) \\ E_k &= E_{\lambda_k}(f_k, \eta_k, M_{k-1}) \ , \end{aligned}$$

$\tau(k), \tau$ defined by (4.2), (4.6), and
λ_{km}, λ_k defined by (4.7), (4.8)
one has:

(i) $f_k \in \mathbb{F}_{k-1}$ (i.e., f_k is admissible for M_{k-1}) and (4.4),

(ii) $Q^{\tau}_{f_k} b \le \eta_k \cdot b$ on $E_k - (E_1 \cup \ldots \cup E_{k-1})$

(iii) $v_k \ge v - \varepsilon_k b$,

(iv) $I(f_k) \ge v - \varepsilon_k b$ on G_k ,

(v) $\lambda_{km}(E_k) \ge \frac{k}{k+1} \lambda_{km}(S)$.

Proof. We make the step from $\{1,\ldots,k-1\}$ to k where k may be 1 and $\{1,\ldots,k-1\}$ may be empty.
By Lemmas 3.8 and 4.1a there is some $f_k \in \mathbf{F}_{k-1}$ such that for $m = 0,\ldots,k-1$

$$\lambda_{km}\left(I(f_k) \ge v_{k-1} - \eta_k^2 b\right) \ge \frac{k}{k+1} \lambda_{km}(S) .$$

For that step we only need that τ is defined on $E_1 \cup \ldots \cup E_{k-1}$ because we consider τ^m only on the set $\left[\xi_i \in E_1 \cup \ldots \cup E_{k-1},\ i = 0,\ldots,m-1\right]$. Define G_k , E_k as above, then by Lemma 4.1b

$$\lambda_k\left(G(f_k, \eta_k^2, M_{k-1}) - E_k\right) = 0$$

hence $\lambda_{km}(\ldots) = 0$, $m = 0,\ldots,k-1$, which implies that

$$\lambda_{km}(E_k) = \lambda_{km}\left(G(f_k, \eta_k^2, M_{k-1})\right) \ge \frac{k}{k+1} \lambda_{km}(S) ,$$

Now (v) is proved. Further (i) follows by construction. Lemma 4.1c implies (ii). Also, by Lemma 4.1d, and the inductive hypothesis

$$v_k \ge v_{k-1} - \eta_k b \ge v - \varepsilon_k b \text{ , i.e. (iii).}$$

and similarly on G_k

$$I(f_k) \ge v_{k-1} - \eta_k b \ge v - \varepsilon_k b \text{ , i.e. (iv) .} \qquad \square$$

Now we can define f according to (4.5). As a consequence, we have from Lemma 4.2 (ii)

$$Q^{\tau}_f b \le \varepsilon \cdot b \text{ on } UE_k . \tag{4.9}$$

If follows from (4.4) and (4.5) that $f = f_k$ on $G_1 \cup \ldots \cup G_k$. Thus starting form an initial state in $E_1 \cup \ldots \cup E_{k-1}$, the

policies f and f_{k-1} use the same actions prior to time τ. Consequently, f_{k-1} can be replaced by f in (4.7). We are using here the fact that $P_{\lambda\delta}(A) = P_{\lambda\delta'}(A)$ if δ and δ' agree prior to τ and $A \in \sigma(\xi_0,\alpha_0,\ldots,\xi_\tau)$.

<u>Lemma 4.3.</u> $P_{\lambda f}\left[\xi_{\tau^m} \notin UE_k,\ldots,\tau^m < \infty\right] = 0$ for all m.

<u>Proof.</u> by induction. For $m = 0$ we have by Lemma 3.10(v):

$$\lambda_{k0}(E_k) \geq \frac{k}{k+1}\,\lambda_{k0}(S)\ ,\ \text{i.e.,}\ \ P_{\lambda f}(\xi_0 \in E_k) \geq \frac{k}{k+1}\ .$$

For the step from $\{0,\ldots,m-1\}$ to m we start from

$$\lambda_{km}(E_k) \geq \frac{k}{k+1}\,\lambda_{km}(S)\ ,\ \text{i.e.,}$$

$$P_{\lambda f}\left[\xi_0 \in E_1 U \ldots UE_{k-1},\ldots,\xi_{\tau^{m-1}} \in E_1 U \ldots UE_{k-1},\xi_{\tau^m} \in E_k\right]$$
$$\geq \frac{k}{k+1}\,P_{\lambda f}\left[\ldots,\xi_{\tau^{m-1}} \in E_1 U \ldots UE_{k-1},\tau^m < \infty\right]$$

Since $[\xi_{\tau^m} \in S] \subset [\tau^m < \infty]$, we get for $k \to \infty$

$$P_{\lambda f}\left[\xi_0 \in UE_k,\ldots,\xi_{\tau^{m-1}} \in UE_k,\xi_{\tau^m} \in UE_k\right] =$$
$$P_{\lambda f}\left[\xi_0 \in UE_k,\ldots,\xi_{\tau^{m-1}} \in UE_k,\tau^m < \infty\right]$$

Now, by the inductive hypothesis,

$$P_{\lambda f}\left[\xi_{\tau^m} \in UE_k\right] = P_{\lambda f}[\tau^m < \infty]\ .\qquad \square$$

<u>Lemma 4.4.</u> There exists a measurable set $E^* \subset UE_k$ such that

$$P_{\lambda f}\left[\xi_{\tau^m} \notin E^*,\tau^m < \infty\right] = 0 \quad \text{for all} \quad m = 0,1,2,\ldots \quad \text{and}$$
$$P_{sf}\left[\xi_{\tau^m} \notin E^*,\tau^m < \infty\right] = 0 \quad \text{for all} \quad m\ ,\ S \in E^*\ .$$

<u>Proof.</u> Write $P = P_{\lambda f}$ and $P_s = P_{sf}$. We will use induction to construct a decreasing sequence $\{E'_n\}$ of measurable sets such that $E'_0 = S$, $E'_1 = UE_k$ and, for $n \geq 1$,

$$P\left[\xi_{\tau^k} \in S - E'_n\right] = 0 \quad \text{for all } k\ ,$$

$$P_s\left[\xi_{\tau^m} \in S - E'_{n-1}\right] = 0 \quad \text{for all} \quad m \quad \text{and all} \quad s \in E'_n .$$

The case $n = 1$ is clear by Lemma 4.3. For the step from n to $n+1$ note that by the inductive hypothesis

$$0 = P\left[\xi_{\tau^k} \in E'_n, \xi_{\tau^{k+m}} \in S - E'_n\right]$$
$$= \int_{E'_n} P_{\xi_{\tau^k}}(ds) P_s\left[\xi_{\tau^m} \in S - E'_n\right] .$$

Now define $E'_{n+1,m} = \left\{s \in E'_n \; ; \; P_s(\xi_{\tau^m} \in S - E'_n) = 0\right\}$ and $E'_{n+1} = \cap_m E'_{n+1,m}$.
Then $P\left[\xi_{\tau^k} \in E'_n - E'_{n+1}\right] = 0$ and by the inductive hypothesis $P\left[\xi_{\tau^k} \in S^\tau - E'_{n+1}\right] = 0$. Now define

$$E^* = \cap E'_n . \quad \square$$

<u>Lemma 4.5.</u> $I(f) \geq v - 3\varepsilon b$ on E^*.

<u>Proof.</u> Set $R = I(f_k)$ on $E_k - (E_1 \cup \ldots \cup E_{k-1})$
$= 0$ on $S - \cup E_k$
Then by Lemma 4.2(iv)

(4.10) $$R \geq v - \varepsilon b \quad \text{on} \quad \cup E_k \supset E^* .$$

On $E^* \cap E_k - (E_1 \cup \ldots \cup E_{k-1})$ we conclude that

$$R = L^\tau_{f_k} I(f_k) \leq L^\tau_{f_k} v = L^\tau_f v$$
$$\leq L^\tau_f (R + \varepsilon b) \qquad \text{by Lemma 4.4 and (4.10)}$$
$$= L^\tau_f R + \varepsilon Q^\tau_f b \qquad \text{by Lemma 3.1}$$
$$\leq L^\tau_f R + \varepsilon^2 b \qquad \text{by (4.9).}$$

Hence

(4.11) $$R \leq L^\tau_f R + \varepsilon^2 b \quad \text{on} \quad E^* .$$

Now one can prove the following assertion by induction

$$(4.12) \qquad R \le L_f^{\tau^n} R + (\textstyle\sum_1^n \varepsilon^{i+1}) b \ , \ Q_f^{\tau^n} b \le \varepsilon^n b \ \text{ on } \ E^* .$$

The case $n = 1$ follows from (4.9) and (4.11). For the step from n to $n+1$, one can show that on E^*

$$\begin{aligned} L_f^{\tau^n} R &\le L_f^{\tau^n}(L_f^{\tau} R + \varepsilon^2 b) && \text{by Lemma 4.4 and (4.11)} \\ &= L_f^{\tau^{n+1}} R + \varepsilon^2 Q_f^{\tau^n} b && \text{by Lemma 3.6} \\ &\le L_f^{\tau^{n+1}} R + \varepsilon^{n+2} b && \text{by the inductive hypothesis} \end{aligned}$$

and

$$\begin{aligned} Q_f^{\tau^{n+1}} b &= Q_f^{\tau^n}(Q_f^{\tau} b) \\ &\le \varepsilon Q_f^{\tau^n} b && \text{by Lemma 4.4 and (4.9)} . \end{aligned}$$

From (4.10) and (4.12) one obtains on E^*

$$\begin{aligned} v - \varepsilon b \le R &\le L_f^{\tau^n} 0 + Q_f^{\tau^n} R + \varepsilon b \quad \text{since} \quad \varepsilon/(1-\varepsilon) \le 1 \\ &\le L_f^{\tau^n} 0 + Q_f^{\tau^n} v + \varepsilon b . \end{aligned}$$

Since $\tau \ge 1$ and hence $\tau^n \ge n$ one can consider that

$$\begin{aligned} Q_f^{\tau^n} v &\le Q_f^{\tau^n} \bar{b} && \text{(Assumption 1)} \\ &\le \bar{b} && \text{(Lemma 2.1)} \end{aligned}$$

and

$$\begin{aligned} \overline{\lim_n}\, Q_f^{\tau^n} v &\le \inf_n \sup_{\sigma \in T, \sigma \ge n} Q_f^{\tau} v \\ &\le V \le b && \text{(Assumption 3)} . \end{aligned}$$

Hence, by Fatou's Lemma and Assumption 3,

$$\begin{aligned} \overline{\lim_n}\, Q_f^{\tau^n} v &= \overline{\lim_n}\, Q_f^{\tau} Q_f^{\tau^{n-1}} v \\ &\le Q_f^{\tau} b \\ &\le \varepsilon b \qquad \text{on } E^* \quad (4.9) \end{aligned}$$

Now we obtain

$$v - \varepsilon b \le \overline{\lim}\, L_f^{\tau^n} 0 + 2\varepsilon b \ .$$

The proof is complete if it is shown that

$$I(f) \ge \overline{\lim}\, L_f^{\tau^n} 0 \quad \text{on} \quad E^* \ . \tag{4.13}$$

This follows from Lemma 3.7b under Assumption 4b or 4c. Under Assumption 4a one has

$$\varepsilon^n b \ge Q_f^{\tau^n} b \ge \inf b \cdot P_{\cdot f}[\tau^n < \infty] \quad \text{on} \quad E^*,$$

hence

$$P_{sf}[\tau^n = \infty] \uparrow 1 \ , \ \text{i.e.} \ \ P_{sf}(\Omega_\tau) = 1 \quad \text{for} \quad s \in E^* \ .$$

Now Lemma 3.7a applies. □

From Lemma 4.4 and Lemma 4.5 we know that

$$P_{\lambda f}\left[\xi_{\tau^m} \in G(3\varepsilon) \,\middle|\, \tau^m < \infty\right] = 1 \quad \text{for all} \quad m \tag{4.14}$$

where

$$G(\eta) = G(f,\eta,M) = \left[I(f) \ge v - \eta b\right] .$$

The epochs τ^m may be called regeneration points as by Kertz (1982). In his terminology we have shown in the last proof that one has terminating regeneration under Assumption 4a.
Now we have to look at epochs between two regeneration points. It is clear from the construction of τ that if the last regeneration has occurred in $E^* \subset UE_k$, which happens with probability one, then the system will be in UG_k up to the next regeneration. We have to distinguish how the next regeneration point is defined, which will depend on the state of the system at the last regeneration point. For that reason we define the disjoint sequence (E_k^*) of measurable sets such that

$$E^* = \sum E_k^* \quad \text{and} \quad \tau = \tau(k) \quad \text{on} \quad E_k^*$$

(i.e. $E_k^* = E^* \cap \left(E_k - (E_1 \cup \ldots \cup E_{k-1})\right)$). Then

$$\tau^{m+1} = \theta_n \circ \tau(k) \quad \text{on} \quad B_{m,n,k} = \left[\tau^m < n < \tau^{m+1} \ , \ \xi_{\tau^m} \in E_k^*\right]$$

From Lemma 4.5 we know that the return earned after the next regeneration is nearly optimal provided that the regeneration occurs in E^*. This will happen almost surely as is shown in the next Lemma.

Lemma 4.6. On $B_{m,n,k}$ one has

$$P_{\xi_n f}\left(\xi_{\tau(k)} \notin E^* \ , \ \tau(k) < \infty\right) = 0 \quad P_{\lambda f} - \text{a.s.}$$

Proof. By Lemma 4.4

$$\begin{aligned} 0 &= P_{\lambda f}\left[B_{m,n,k} \ , \ \xi_{\tau^{m+1}} \in S - E^*\right] \\ &= P_{\lambda f}\left[B_{m,n,k} \ , \ \xi_{\tau(k)} \circ \theta_n \in S - E^*\right] \\ &= \int_{B_{m,n,k}} P_{\xi_n f}\left[\xi_{\tau(k)} \in S - E^*\right] \end{aligned}$$ □

Furthermore we know from Lemma 4.2(iv) that the return earned after the present epoch n is nearly optimal under a policy which agrees with f up to the next regeneration. These facts will be combined in the next Lemma.

Lemma 4.7. On $B_{m,n,k}$ one has

$$\xi_n \in G(4\varepsilon) \quad P_{\lambda f} - \text{a.s.}$$

Proof. On $B_{m,n,k}$ one has $\xi_n \in G_k$. Hence by Lemma 4.2(iv)

$$\begin{aligned} v(\xi_n) - \varepsilon b(\xi_n) &\le I(f_k, \xi_n) \\ &= L_{f_k}^{\tau(k)} I(f_k)(\xi_n) = L_f^{\tau(k)} I(f_k)(\xi_n) \\ &\le L_f^{\tau(k)} v(\xi_n) \\ &\le L_f^{\tau(k)} (I(f) + 3\varepsilon b)(\xi_n) \quad P_\lambda - \text{a.s.} \quad \text{by Lemma 4.6} \\ &= L_f^{\tau(k)} I(f)(\xi_n) + 3\varepsilon Q_f^{\tau(k)} b(\xi_n) \\ &\le I(f, \xi_n) + 3\varepsilon b(\xi_n) \quad \text{by Lemma 2.1 .} \end{aligned}$$ □

From (4.14) and Lemma 4.7 we can conclude

$$\xi_n \in G(4\varepsilon) \quad \text{for all} \quad n \quad P_{\lambda f} - \text{a.s.} \tag{4.15}$$

and Theorem 1 is proved.
Finally we will prove Theorem 2. The proof is adapted from Dubins and Sudderth (1979). We assume that Condition A holds and choose $\lambda = \mu$ in the preceding analysis. One knows that v is upper semianalytic. Combining the Propositions 7.48 and 7.50 of Bertsekas, Shreve (1978), one obtains the existence of a stationary policy $f' \in F_u$ which always uses "εb-conserving actions", i.e.

$$L_{f'}^1 v + \varepsilon b \geq v \ .$$

Note that, by Assumption 5a, the supremum in the optimality equation is attained on $S_b = [b = 0]$. (The definition of f' on S_b is clear if (2.0) holds.)
Let

$$\begin{aligned} f^* &= f \quad \text{on} \quad E^* \\ &= f' \text{ on} \quad S - E^* \ . \end{aligned}$$

From Assumption A and Lemma 4.4 (with $m = 0$) one knows that

$$\xi_n \in E^* \quad n = 1,2,\ldots \quad P_{sf^*}\text{- a.s.} \quad \text{for all} \quad s \in S \ . \tag{4.16}$$

Hence $I(f^*) = I(f)$ on E^* . Further on $S - E^*$

$$\begin{aligned} I(f^*) &= L_{f'}^1 I(f) \\ &\geq L_{f'}^1 (v - 3\varepsilon b) \qquad \text{by (4.16) and Lemma 4.5} \\ &= L_{f'}^1 v - 3\varepsilon Q_{f'}^1 b \\ &\geq v - \varepsilon b - 3\varepsilon b = v - 4\varepsilon b \ . \end{aligned}$$

Thus we can conclude

$$I(f^*) \geq v - 4\varepsilon b \quad \text{on} \quad S \ . \tag{4.17}$$

Acknowledgement. The authors are grateful to Dr. van Dawen for several useful discussions on the topic of this paper.

References

Barbosa - Dantas, C.A. (1966). The existence of stationary optimal plans. Dissertation. Univ. of California, Berkeley

Bertsekas, D.P., Shreve, S.E. (1978). Stochastic Optimal Control Academic Press New York

Blackwell, D. (1965). Positive dynamic programming. Proc. Fifth Berkeley Sym. Math. Statist. Prob. 1, 415-418

Blackwell, D. (1976). The stochastic processes of Borel gambling and dynamic programming. Annals of Statistics 4, 370-374

Blackwell, D., Freedman, D., and Orkin, M. (1974). The optimal reward operator in dynamic programming. Annals of Probability 2, 926-941

Bodewig, H.-H. (1979). Über dynamische Optimierung mit endlich additiven Maßen. Master thesis, Univ. Bonn

Brown, L.D., Purves, R. (1973). Measurable selections of extrema. Ann. Statist. 1, 902-912

Dawen, van R. (1983). Stationäre Politiken in stochastischen Entscheidungsproblemen. Dissertation, Univ. of Bonn, Dept. of Appl. Math. (To be published in English)

Dawen, van R. (1984). On stationary strategies in positive stochastic 1 and 2 person games with general state space. ZAMM

Dawen, van R., Schäl, M. (1983a). On the existence of stationary optimal policies in Markov decision models. ZAMM 63, T403-T404

Dubins, L.E., Savage, L.J. (1965). How to gamble if you must. Mc Graw-Hill

Dubins, L.E., Sudderth, W. (1977a). Persistently ε-optimal strategies. Math. of Op. Res. 2, 125-134

Dubins, L.E., Sudderth, W. (1977b). Countable additive gambling and optimal stopping. Z.f. Wahrscheinlichkeitstheorie 41, 59-73

Dubins, L.E., Sudderth, W. (1979). On stationary strategies for absolutely continuous houses. Ann. Prob. 7, 461-476

Engelbert, A., and Engelbert, H.J. (1979). Optimal stopping and almost sure convergence of random sequences. Z.f. Wahrscheinlichkeitstheorie 48, 309-325

Federgruen, A., Hordijk, A., Tijms, H.C. (1979). Denumerable state semi-Markov decision processes with unbounded costs. Average cost criterion. Stoch. Proc. Appl. 9, 223-235

Frid, E.B. (1976). On a problem of D. Blackwell from the theory of dynamic programming, Theor. Probability. Appl. 15, 719-722

Hinderer, K. (1970). Foundations of non-stationary dynamic programming with discrete time-parameter. Lecture Notes Operations Research and Math. Systems 33, Springer-Verlag

Kertz, R.P. (1982). Renewal plans and persistent optimality in countable additive gambling. Math. of Op. Res. 7, 361-382

Schäl, M. (1975). Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Z.f. Wahrscheinlichkeitstheorie 32, 179-196

Schäl, M. (1983). Stationary policies in dynamic programming models under compactness assumptions. Math. of Op. Res. 8, 366-372

Strauch, R.E. (1966). Negative dynamic programming. Ann. Math. Statist. 37, 871-890

Sudderth, W.D. (1969). On the existence of good stationary strategies. Trans. Am. Math. Soc. 135, 399-414

Sudderth, W.D. (1971). A 'Faton equation' for randomly stopped variables. Annals of Math. Statist. 42, 2143-2146

Wal, van der J. (1981). On uniformly nearly-optimal stationary strategies. Eindhoven Univ. of Techn., Dept. of Math., Memorandum COSOR 81-14

Wal, van der J. (1983). On uniformly nearly-optimal Markov strategies. Operations Research Proceedings 1982, 461-467